# ICA-BASED MAP ALGORITHM FOR SPEECH SIGNAL ENHANCEMENT

*Xin Zou*[1], *Peter Jančovič*[1], *Ju Liu*[2] *and Münevver Köküer*[1]

[1] Electronic, Electrical & Computer Engineering, University of Birmingham, Birmingham, UK
[2] School of Information Science and Engineering, Shandong University, Shandong, China

{xxz391, p.jancovic, m.kokuer }@bham.ac.uk, juliu@sdu.edu.cn

## ABSTRACT

This paper proposes a novel MAP denoising algorithm that uses the ICA transformation and provide a derivation demonstrating the type of situations in which the use of ICA transformation is expected to achieve best results. We also propose an employment of Generalized Gaussian Model (GGM) for modelling the speech and noise distributions. The performance of the proposed speech enhancement algorithm is compared with the Wiener filtering and sparse code shrinkage method. The experiments are focused on speech signal corrupted by a non-Gaussian noise. The experimental results show that the proposed algorithm achieves significantly better performance than both the Wiener filtering and the sparse code shrinkage method.

***Index Terms—*** Speech enhancement, independent component analysis, maximum a-posteriori estimation, generalized Gaussian model, non-Gaussian noise

## 1. INTRODUCTION

Speech enhancement algorithms aim at improving the quality of noise-corrupted speech signals by removing the corrupting noise. This has attracted a great deal of attention over past several decades.

A traditional method to enhance speech signal degraded by an additive stationary noise is spectral subtraction [1]. The Wiener filtering, assuming both the speech and noise signals to have Gaussian distribution, has widely been used for speech enhancement, e.g., [2]. In recent studies, the use of MAP algorithm has been proposed, e.g., [3] [4] [5]. The estimation is usually carried out in a linear transformation domain. The use of various transform-domains has been investigated, for instance, the KLT (PCA) [3], wavelet transform [4], or ICA transform [5]. It has also been reported that a non-Gaussian modelling of speech signals may result in a better performance. The use of Laplace distribution was proposed in [3] [4], while authors in [5] used two super Gaussian distribution models. Several open topics that need further investigation include: finding an efficient linear transformation for a specific noisy conditions, dealing with a non-Gaussian noise corruption, and a more flexible distribution modelling.

In this paper, we attempt to address some of the above questions. We propose a novel MAP denoising algorithm that uses the ICA transformation and provide a derivation that demonstrates the type of situations (i.e. speech/noise being Gaussian/non-Gaussian) in which the use of ICA transformation is expected to achieve best results. We also propose an employment of Generalized Gaussian Model (GGM), originally introduced in [6], as a flexible model for modelling a wide class of non-Gaussian distributions. The performance of the proposed speech enhancement algorithm is compared with the Wiener filtering and sparse code shrinkage method. In our experiments, we focus on speech corrupted by a non-Gaussian noise. The experimental results show that the proposed algorithm achieves significantly better performance than both the Wiener filtering and the sparse code shrinkage method.

The paper is organised as follows. The MAP algorithm based on the ICA is presented in Section 2. The denoising capability analysis of the proposed algorithm is discussed in Section 3. The GGM model and the related parameter estimation is introduced in Section 4. Section 5 presents simulation results and conclusions are given in Section 6.

## 2. MAXIMUM A POSTERIORI ESTIMATION

We consider scalar random variables. Denote by $x$ the original speech signal, and by $v$ some additive noise. Assume that we have observed the random variable $y$ which is the noisy version of signal $x$, i.e.,

$$y = x + v \tag{1}$$

Our task is to estimate $x$ from the observed noisy signal $y$ by means of $\hat{x} = g(y)$ . This estimation problem may be solved by using the MAP algorithm.

Let us denote an $N$-dimensional vector of samples of noisy speech signal $y(t)$ at time $t$ by $\mathbf{y}(t) = [y(t), \cdots, y(t - N + 1)]^T$, where $(.)^T$ denotes the transpose operation. The entire noisy signal gives a set of $N$-dimensional observation vectors $\{\mathbf{y}(1), \cdots, \mathbf{y}(T)\}$. Corresponding notation for clean speech signal. The MAP-based estimation of

$\{\hat{\mathbf{x}}(1), \cdots, \hat{\mathbf{x}}(T)\}$ for large $T \to \infty$ can be obtained by

$$\begin{aligned} & \{\hat{\mathbf{x}}(1), \cdots, \hat{\mathbf{x}}(T)\} \\ = \quad & \arg\max_{\mathbf{x}} \sum_t (\ln p_v(\mathbf{y}(t) - \mathbf{x}(t)) + \ln p_x(\mathbf{x}(t))) \\ \Leftrightarrow \quad & \arg\max_x (E\{\ln p_v(\mathbf{y}(t) - \mathbf{x}(t))\} + E\{\ln p_x(\mathbf{x}(t))\}) \end{aligned} \tag{2}$$

Our analysis considers a linear transformation framework. For convenience, the time index $t$ will be omitted in the following derivations. It has been shown in [7] that the maximization in Eq. 2 can be achieved by using ICA-based independent components. To derive our result, we will use the relationship between the density of a random variable $a$ and its linearly transformed version $b = Wa$, i.e. $p_a(\mathbf{a}) = p_b(W\mathbf{a})|\det(W)|$, where det denote determinant of matrix.

Considering estimation of a single frame, the expectation operator can be removed, and then the ICA-based MAP algorithm becomes in the form of

$$\begin{aligned} \hat{\mathbf{x}} \quad \leftarrow \quad & \arg\max_x (\ln p(W^v(\mathbf{y} - \mathbf{x})) + \ln p(W^x \mathbf{x}) \\ & + \ln|\det(W^x)||\det(W^v)|) \end{aligned} \tag{3}$$

where $W^v$ and $W^x$ are ICA matrices of noise and clean signal respectively. In the ICA framework, these matrices depend only on the given noise and clean signal, and thus can be considered as constant. Let us denote the right term of Eq. 3 by $L$. The $n$th sample estimate of the signal, $\hat{x}_n$, can be updated by using gradient method

$$\begin{aligned} x_n \quad & \leftarrow x_n + \lambda \frac{\partial L}{\partial x_n} \\ = \quad & x_n + \lambda \sum_{k=1}^{N} [-f_v'(W_{k\cdot}^v(\mathbf{y} - \mathbf{x}))w_{kn}^v + f_x'(W_{k\cdot}^x \mathbf{x})w_{kn}^x] \end{aligned} \tag{4}$$

where $\lambda$ is the step size, $f = \ln p$, $f'$ is the partial derivative of $f$ with respect to $\mathbf{x}$ and $w_{kn} = W(k, n)$, $W_{k\cdot} = W(k, :)$.

Assuming the noise is Gaussian, it can be shown that the Eq. 3 can lead to the best linear estimator, i.e. Wiener filter, or sparse code shrinkage method.

## 3. DENOISING CAPABILITY ANALYSIS

In this section, we analyse the denoising capability of the MAP estimation given by Eq. 2 for different types of random variables. In Bayesian estimation, the estimators of some variables are achieved by minimizing a conditional risk $R$ which is given by a cost function $C(\hat{\mathbf{x}}, \mathbf{x})$ of estimating the true value of $\mathbf{x}$ as $\hat{\mathbf{x}}$:

$$\begin{aligned} R_{\mathbf{x}}(\hat{\mathbf{x}}) \quad = \quad & E\{C(\hat{\mathbf{x}}, \mathbf{x})\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C(\hat{\mathbf{x}}, \mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\ = \quad & \int_{-\infty}^{\infty} [\int_{-\infty}^{\infty} C(\hat{\mathbf{x}}, \mathbf{x}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x}] p(\mathbf{y}) d\mathbf{y} \end{aligned} \tag{5}$$

where $\mathbf{y}$ is the observed data and $\mathbf{x}$ is the true value of data hidden in $\mathbf{y}$.

Often we do not know the a priori probability density of an observation $\mathbf{y}$ and we simply assign it to be a uniform distribution in our following derivation. As the Bayesian risk is always positive, thus the minimization of $R_{\mathbf{x}}(\hat{\mathbf{x}})$ is obtained by selecting $\hat{\mathbf{x}}$ with which for the given $\mathbf{y}$ the term in bracket in Eq. 5 is minimum. Several forms of cost function can be chosen, which all depend on the problem to be solved. Suppose that in a given estimation problem we are not able to assign a particular cost function $C(\hat{\mathbf{x}}, \mathbf{x})$, then a natural choice is a uniform cost function equal to 0 between some certain small values $\pm\epsilon/2$ and uniform value outside that. Then by Bayes theorem the risk function $R_{\mathbf{x}}(\hat{\mathbf{x}})$ can be expressed by the posteriori probability as

$$R_{\mathbf{x}}(\hat{\mathbf{x}}) = E\{\frac{1}{\epsilon} - p(\hat{\mathbf{x}}|\mathbf{y})\} = E\{\frac{1}{\epsilon} - \frac{p(\mathbf{y}|\hat{\mathbf{x}})p(\hat{\mathbf{x}})}{p(\mathbf{y})}\} \tag{6}$$

where $p(\mathbf{y})$ can be treated as a uniform probability distribution, i.e. $p(\mathbf{y}) = 1/T$. This equation shows that the minimum of $R_{\mathbf{x}}(\hat{\mathbf{x}})$ equals to the maximization of $E\{p(\mathbf{y}|\hat{\mathbf{x}})p(\hat{\mathbf{x}})\}$. Without affecting the convergence point, we consider the pdf to be compressed by a logarithm function. Without causing any confusion, we just use $\mathbf{x}$ to replace the estimator $\hat{\mathbf{x}}$ in our following derivation.

As we discussed in Section 2, the maximization of log likelihood can be solved within ICA framework. For the term $E\{\ln p(\mathbf{x})\}$, some simple manipulations [8] yields

$$E\{\ln p(\mathbf{x})\} = -H(\mathbf{x}) - KL(\mathbf{s} \parallel \mathbf{a}) \tag{7}$$

where $KL(\cdot \parallel \cdot)$ and $H(\cdot)$ denote Kullback-Leibler divergence and entropy respectively. The $\mathbf{a}$ is the linear transformation of $\mathbf{x}$ and $\mathbf{s}$ is the independent sources of vector $\mathbf{x}$, which can be modelled as

$$\mathbf{x} = A\mathbf{s} \tag{8}$$

where $A$ is the invertible mixing matrix. $H(\mathbf{x})$ is the entropy of the true distribution of $\mathbf{x}$, i.e. the entropy of vector $A\mathbf{s}$. By the linear transformation of entropy given in [9], we have

$$H(\mathbf{x}) = H(\mathbf{s}) + \ln|\det(A)| \tag{9}$$

In ICA blind framework, the value of $\ln|\det(A)|$ is equal to $-\ln|\det(U)|$, where $U$ is whitening matrix of $\mathbf{x}$. As it has been shown in [10], the maximum log likelihood can be achieved by the minimization of $KL(\mathbf{s} \parallel \mathbf{a})$, which would result in zero, as the KL divergence is being non-negative. As such the maximum log likelihood of $p(\mathbf{x})$ will be

$$E\{\ln p(\mathbf{x})\}_{\max} = -H(\mathbf{s}) + \ln|\det(U)| \tag{10}$$

The same analysis as above can also be applied for the term $E\{\ln p_v(\mathbf{y} - \mathbf{x})\}$, yielding a similar expression as in Eq. 10.

Let us denote the independent components of noise and original clean signal by $s_v$ and $s_x$, respectively, and the whitening matrix of noise and original clean signal by $U_v$ and $U_x$, respectively. To evaluate the denoising capability of the proposed algorithm for signals/noises of various statistical properties, the variance of each given signal and each given noise is assumed to be fixed (i.e. SNR is fixed). Since the values of $\det(U_v)$ and $\det(U_x)$ only correspond to the variance of a given noise and signal, respectively, they can be considered to be constants. The maximum value of the term $E\{\ln(p(\mathbf{y}|\mathbf{x})p(\mathbf{x}))\}$ can then be approximated as

$$\begin{aligned} & E\{\ln(p(\mathbf{y}|\mathbf{x})p(\mathbf{x}))\}_{max} \\ = \; & -H(\mathbf{s_v}) + \ln|\det(U_v)| - H(\mathbf{s_x}) + \ln|\det(U_x)| \\ = \; & -H(\mathbf{s_v}) - H(\mathbf{s_x}) + C \end{aligned} \tag{11}$$

where $C = \ln|\det(U_v)| + \ln|\det(U_x)|$ is a constant.

As the entropy of a non-Gaussian variable can be approximately related to negative of its kurtosis [9], the value of entropy terms $H(\cdot)$ in Eq. 11 will be smaller as the variable $\mathbf{s}$ is more non-Gaussian (i.e. higher kurtosis), and will have maximum value for a Gaussian variable. Then, considering Eq. 6 and Eq. 11, the minimum of Bayes risk denoted by $R_{x,v}$ will be in the order of

$$R_{(x_{ng},v_{ng})} < R_{(x_{ng},v_g)} < R_{(x_g,v_g)} \tag{12}$$

where the subscripts $ng$ and $g$ denote non-Gaussian and Gaussian distribution, respectively. As can be seen, the denoising capability improves with increasing the non-Gaussianity of signal.

Note that when both the signal and noise are Gaussian, the MAP estimator lead to the best linear estimator, i.e. Wiener filter denoising. This means that allowing the nonlinearity in the estimation would not improve the performance when both the signal and noise are Gaussian. On the other side, the proposed algorithm can achieve best results in the case of a non-Gaussian signal corrupted by a non-Gaussian noise.

## 4. THE PROBABILITY DENSITY MODEL

### 4.1. Generalized Gaussian Model

Another important issue of MAP is to model the density function for the given random variables. The model should be a good fit to various degrees of non-Gaussian (while still being computationally feasible). In this paper we propose to employ the Generalized Gaussian Model (GGM). Box and Tiao [6] expressed GGM in the following general form

$$p(z|\mu,\delta,\beta) = \frac{\omega(\beta)}{\delta} exp[-c(\beta)|\frac{z-\mu}{\delta}|^{2/(1+\beta)}] \tag{13}$$

where

$$c(\beta) = [\frac{\Gamma[3(1+\beta)/2]}{\Gamma[(1+\beta)/2]}]^{1/(1+\beta)} \tag{14}$$

$$\omega(\beta) = \frac{\Gamma[3(1+\beta)/2]^{1/2}}{(1+\beta)\Gamma[(1+\beta)/2]^{3/2}} \tag{15}$$

where $\Gamma$ is gamma function.

The $\mu$ and $\delta$ denote the mean and standard deviation of the data, respectively. The parameter $\beta$ controls the deviation of the distribution from Gaussian; by varying $\beta$, the Eq. 13 can describe Gaussian, sub-Gaussian, super-Gaussian distributions. For instance, when $\beta = 0$ the distribution is Gaussian and when $\beta = 1$ it is a Laplacian. As $\beta \to -1$, the distribution becomes uniform, as $\beta \to \infty$, the distribution is a delta function.

For zero mean and unit variance variable $z$, the function $f'$ in Eq. 4 can be expressed as

$$f'(z) = \frac{-2c(\beta)}{1+\beta}|z|^{2/(1+\beta)-1} \tag{16}$$

### 4.2. Parameter estimation

In our case, the signal is assumed to be zero mean and unit variance, the problem then resorts to the estimation of the value of $\beta$. By MAP method, the parameter $\beta$ can be obtained by maximization of the posteriori density function given the training signal $x = \{x(1), \cdots, x(T)\}$, i.e.,

$$\beta = \arg\max_{\beta} p(\beta|x) \Leftrightarrow \arg\max_{\beta} p(x|\beta)p(\beta) \tag{17}$$

where the data likelihood is:

$$p(x|\beta) = \prod_t \omega(\beta) exp[-c(\beta)|x(t)|^{2/(1+\beta)}] \tag{18}$$

and $p(\beta)$ defines the prior distribution of $\beta$, which is suggested in [11] to be modelled by Gamma function, i.e. $p(\beta) \backsim Gamma(\beta|c_1, c_2)$ , where $c_1$ and $c_2$ are constant. Choosing the values $c_1 = c_2 = 2$ can give a broad distribution with 95% densities range of $\beta \in [-0.5, 10.5]$ [11].

## 5. EXPERIMENTAL RESULTS

Since the MAP based Gaussian noise signal enhancement, i.e. sparse code shrinkage and best linear estimation (Wiener filter) have been well researched, our evaluation focuses on enhancement of signals corrupted by a non-Gaussian noise. The proposed algorithm is evaluated with speech signals from the TIMIT database, separately for each gender. The training set contained 30 sentences for each gender, which were randomly selected from the DR1 subset, and testing set contained five sentences of speaker fdaw0 (female) and mcpm0 (male); sentences from testing speakers were not used in the training set. Noisy speech signals are created by adding the Railway and Pub noise from the Noisex92 database to the clean speech signal at SNR=0dB. Half of the noise signals is used to train the ICA basis functions, while the other half is used to corrupt the test signals. All of the signals are sampled at 8 kHz.

During training, the signals are segmented into frames of 64 samples by rectangular window, and the fast-ICA algorithm [9] is applied to extract ICA basis functions from the training data. The results obtained by the proposed algorithm, compared with standard Wiener filter and sparse code shrinkage method, are presented in Table 1. It can be seen that the proposed algorithm significantly improves the SNR for both genders in both noisy conditions. The time-domain waveforms of the original clean signal, noise corrupted signal, and enhanced signal by the proposed algorithm are illustrated in Figure 1.

**Table 1**. *The comparison of the proposed ICA-based MAP algorithm, standard Wiener filter and sparse code shrinkage.*

| | Enhancement algorithm | Enhanced SNR Pub | Enhanced SNR Railway |
|---|---|---|---|
| Male | Proposed method | 6.9 | 7.0 |
| | Wiener filter | 0.8 | 2.6 |
| | Sparse code shrinkage | 1.3 | 1.1 |
| Female | Proposed method | 5.9 | 10.8 |
| | Wiener filter | 1.9 | 3.7 |
| | Sparse code shrinkage | 2.8 | 2.5 |

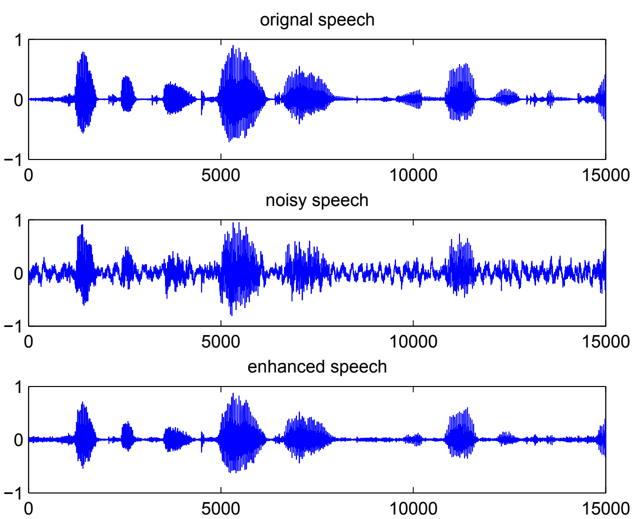


**Fig. 1**. *Speech signal enhancement result for female speech corrupted by Railway noise at 0dB.*

## 6. CONCLUSION

In this paper, we presented the ICA based MAP speech signal enhancement method. The denoising capabilities of the proposed algorithm were analysed considering for both speech and noise signals both Gaussian and non-Gaussian distribution. The analysis result shows that the proposed method is most efficient for non-Gaussian signals corrupted by a non-Gaussian noise. The generalized Gaussian model was proposed to be used in modelling the density function in the MAP. The experiments were conducted for enhancement of speech corrupted by additive non-Gaussian noises. The obtained results showed significant performance improvement achieved by the proposed algorithm in comparison to Wiener filtering and sparse code shrinkage.

This work was supported by UK EPSRC grant EP/D033659/1. And partially supported by Program for New Century Excellent Talents in University, China (NCET-05-0582).

## 7. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic, Speech, and Signal Proc.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] I.Y. Soon and S.N. Koh, "Low distortion speech enhancement," *Proc. Inst. Elec. Eng.*, vol. 147, no. 3, pp. 247–253, 2000.

[3] S. Gazor and W. Zhang, "Speech enhancement employing laplacian-gaussian mixture," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 896–904, 2005.

[4] H. Xie, L.E. Pierce, and F.T. Ulaby, "Sar speckle reduction using wavelet denoising and markov random field modeling," *IEEE Transactions on geoscience and remote sensing*, vol. 40, no. 10, pp. 2196–2212, 2002.

[5] A. Hyvärinen, *Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation*, vol. 11, Neural Computation, 1999.

[6] G. Box and G. Tiao, *Bayesian inference in statistical analysis*, John Wiley and Sons, 1973.

[7] X. Zou, P. Jančovič, and J. Liu, "The efficience of ica-based reprentation: application to speech feature extraction for noise robust speaker recognition," *EUSIPCO*, 2006.

[8] J.F. Cardoso, "Informax and maximum likelihood for blind source separation," *IEEE signal processing letters*, vol. 4, no. 4, 1997.

[9] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons, Inc., 2001.

[10] J.F. Cardoso, "High-order constrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.

[11] T. W. Lee and G. J. Lewicki, *The generalized Gaussian mixture model using ICA*, In Proc. Int. Workshop on Independent Component Analysis, 2000.